(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 1 022 725 A1

(12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
26.07.2000 Bulletin 2000/30

(51) Int Cl.[7]: **G10L 17/00**, G10L 15/06

(21) Application number: **99100951.5**

(22) Date of filing: **20.01.1999**

(84) Designated Contracting States:
AT BE CH CY DE DK ES FI FR GB GR IE IT LI LU MC NL PT SE
Designated Extension States:
AL LT LV MK RO SI

(71) Applicant: **Sony International (Europe) GmbH**
10785 Berlin (DE)

(72) Inventors:
- **Goronzy, Silke c/o Sony Int. (Europe) GmbH**
  70736 Fellbach (DE)
- **Buchner, Peter c/o Sony Int. (Europe) GmbH**
  70736 Fellbach (DE)
- **Kompe, Ralf c/o Sony Int. (Europe) GmbH**
  70736 Fellbach (DE)

(74) Representative:
MÜLLER & HOFFMANN Patentanwälte
Innere Wiener Strasse 17
81667 München (DE)

(54) Selection of acoustic models using speaker verification

(57) Usually in a speaker adaptive system, every time a change in speaker occurs, he/she has to chose which of the available model sets to use. E.g. the SI model set, if it is the first time he/she is using the system or if a model set already adapted to him, if he/she used it before. If adapted model sets are not stored at all, the adaptation process starts over and over again using the SI models, if the same speaker uses the system repeatedly. According to the invention a change in speaker is be detected automatically. Furtheron, the system identifies the speaker and if he/she had used the system before and a speaker adapted model to him/her is already available. If this is the case, this model set will be taken for further recognition and adaptation.

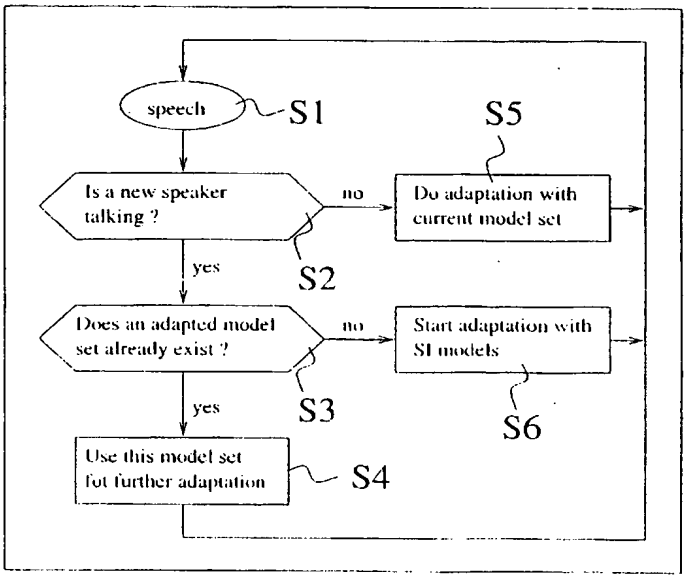

Fig. 2

EP 1 022 725 A1

Description

[0001] This invention is related to a method and a device to perform automatic speech recognition, in particular to a method and a device to increase the recognition rate in speech recognition systems that are used by different users.

[0002] State of the art speech recognizers consist of a set of statistical distribustions modeling the acoustic properties of certain speech segments. These acoustic properties are encoded in feature vectors. As an example, one Gaussian distribution can be taken for each phoneme. These distributions are attached to states. A (stochastic) state transition network (usually hidden Markov models) defines the probabilities for sequences of states and sequences of feature vectors. Passing a state consumes one feature vector covering a frame of e.g. 10 ms of the speech signal.

[0003] The stochastic parameters of such a recognizer are trained using a large amount of speech data either from a single speaker yielding a speaker dependent (SD) system or from many speakers yielding a speaker independent (SI) system.

[0004] Speaker adaptation (SA) is a widely used method to increase recognition rates of SI systems. State of the art speaker dependent systems yield much higher recognition rates than speaker independent systems. However, for many applications, it is not feasible to gather enough data from a single speaker to train the system. In case of a consumer device this might even not be wanted. To overcome this mismatch in recognition rates, speaker adaptation algorithms are widely used in order to achieve recognition rates that come close to speaker dependent systems, but only use a fraction of speaker dependent data compared to speaker dependent ones. These systems initially take speaker independent models that are then adapted so as to better match the speakers acoustics.

[0005] Usually, the adaptation is performed in supervised mode. That is the spoken words are known and the recognizer is forced to recognize them. Herewith a time alignment of the segment-specific distributions is achieved. The mismatch between the actual feature vectors and the parameters of the corresponding distribution builds the basis for the adaptation. The supervised adaptation requires an adaptation session to be done with every new speaker before he/ she can actually use the recognizer.

[0006] Usually, the speaker adaptation techniques modify the parameters of the hidden Markov models so that they better match the acoustic characteristics of new speakers. Normally, in batch or off-line adaptation a speaker has to read a pre-defined text before he/she can use the system for recognition, which is then processed to do the adaptation. Once this is finished the system can be used for recognition. This mode is also called supervised adaptation, since the text was known to the system and a forced alignment of the corresponding speech signal to the models corresponding to the text is performed and used for adaptation.

[0007] However, an unsupervised or on-line method is better suited for most kinds of consumer devices. In this case, adaptation takes place while the system is in use. The recognized utterance is used for adaptation and the modified models are used for recognizing the next utterance and so on. In this case the spoken text is not known to the system, but the word(s) that were recognized are taken instead.

[0008] An adapatation of the speaker adapted model set can be repreatedly performed to further improve the performance of specific speakers. There are several existing methods for speaker adaptation, e.g. maximum a posteriori adaptation (MAP) or maximum likelihood linear regression (MLLR) adaptation.

[0009] However, these speaker adaptive speech recognition systems, especially systems working with unsupervised adaptation, are always adaptated to one speaker only. Therefore, if the speaker changes, adaptation has to be restarted (using the SI models) for this new speaker before he/she can use the system with an improved recognition rate.

[0010] Speaker adaptation techniques are widely used in many kinds of speech recognition systems, e.g. dictation systems. In some of these systems it is possible to store the speaker adapted models, so that different speakers can use the system with different speaker adapted models. But each time it has to be specified manually which of the adapted models to use.

[0011] On the other hand, it is known that speaker verification and identification techniques are used for access control of e.g. buildings or systems.

[0012] Therefore, it is the object underlying the present invention to propose a method and a device for speaker adaptation that overcomes the problems described above.

[0013] The inventive method is defined in independent claim 1 and the inventive device is defined in independent claim 5. Preferred embodiments thereof are respectively defined in the respective following dependent claims.

[0014] As mentioned above, according to the prior art adaptation has to be restarted using the speaker independent (SI) models again if there is change in speaker.

[0015] When talking about a home or car environment there will be a change in speaker quite often, but it will be a more or less fixed set of speakers, e.g. the members of a family. So it is not very reasonable to start adaptation all over again every time one of the speakers starts using the system and discard all previous adaptation to specific speakers.

[0016] According to the present invention, on the other hand, the system recognizes the speaker, and if adaptation has already been conducted for that speaker. the models already existing will be used for further adaptation. Speaker verification techniques are used for recognizing who is speaking.

[0017] According to the present invention this change in speaker is detected automatically. Therefore, in a networked system that is mainly used by the same persons, but with a frequent change between them, the speech recognition system according to the present invention does not restart the adaptation to a different speaker every time the speaker changes, but it first checks the identity of the speaker, so that the system can switch to an adapted model set for this particular speaker, if it exists. In this case, said model set is stored and used for recognition and further adaptation. Together with the speaker adapted model set, the statistical hyper parameters necessary for the adaptation are stored so that the adaptation can continue and does not have to be restarted when the same speaker uses the system again. Such hyper parameters could e.g. be weights that determine the adaptation speed to adapt a certain speaker adapted model set to the corresponding speaker. If no model set exists for this particular speaker, a new one will be built using adaptation starting with the SI models.

[0018] The method and device according to the present invention will be better understood from the following detailed description of an exemplary embodiment thereof taken in conjunction with the appended drawings, wherein:

Figure 1 shows a speech recognition system according to the present invention using speaker adaptation and automatic identification of the speaker; and

Figure 2 shows the verification and adaptation procedure performed according to the present invention.

[0019] Figure 1 only shows the part of the automatic speech recognition system according to the present invention that is used for speaker adaptation and automatic identification of the speaker.

[0020] The analogue speech signal generated by a microphone 1 is converted into a digital signal in an A/D conversion stage 2 before a feature extraction is performed by a feature extraction module 3 to obtain a feature vector, e.g. every 10 ms. This feature vector is fed into a verification module 4 and a recognition module 5. In the verification module 4 an automatic identification of the speaker is performed, as described above. In the recognition module 5 recognition of the spoken utterance is performed on basis of the extracted feature vectors and a set of HMM models. The recognition module 5 also feeds the recognition result to an adaptation module 6 that can adapt a certain HMM model set to a certain speaker.

[0021] The HMM model set to be accessed or adapted by the recognition module 5 or the adaptation module 6 is selected by the verification module 4 from a speaker independent model set or one of several sets of speaker adapted model sets that are respectively adapted to different individual speakers. These different model sets are stored in storages 7, 8, 9 and 10 and selected via a switch 11 that has its fixed terminal connected to the recognition module 5 and the adaptation module 6 and the movable terminal dependent on a control signal that is received from the verification module 4 to one of the model sets described before.

[0022] It is also possible that the speaker adapted model sets are not adaptated to individual speakers, but to individual groups of speakers, such as Germans, British people, Germans speaking English, American people and so on or people speaking different dialects. These groups can also be identified automatically according to well known language or dialect identification algorithms working directly on the speech signal.

[0023] Of course, instead of the switch 11 a different solution having the same function can be selected.

[0024] Figure 2 shows the verification and adaptation procedure performed in the recognition system according to the present invention. In a first step S1 a spoken utterance of a user is received, A/D converted and further processed to extract the feature vectors. Thereafter, it is checked in a step S2 whether a new speaker is talking or not. If a new speaker is talking, it is checked in step S3 whether an adapted model set already exists for this speaker or not. If an adapted model set already exists this model set is used for further adaptation in a step S4, whereafter the next spoken utterance is processed in step S1 and the whole procedure is repeated therewith.

[0025] If no adapted model set exists in step S3, adaptation with the speaker independent model is started in step S6 and a new model set (speaker adapted) is added to the system, whereafter the next utterance is processed in step S1 and the whole process will be repeated with his next utterance. If it is determined in step S2 that no new speaker is talking, the adaptation will be done with the current model set in step S5, whereafter the next spoken utterance is processed in step S1 and the whole procedure will be repeated with this next utterance.

Claims

1. Method to perform an automatic speech recognition, **characterized in that**

   a change of the speaker is detected automatically;
   a speaker gets identified; and
   an individual model set adapted to the identified speaker is used for the speech recognition procedure, if it is available, otherwise such an individual speaker adapted model is newly generated for said speaker.

2. Method according to claim 1, **characterized in that** an individual speaker adapted model set is gener-

ated on basis of a speaker independent model.

3. Method according to claim 1 or 2, **characterized in that** an individual speaker adapted model set is adaptated on basis of utterances of the corresponding speaker.

4. Method according to claim 1, 2 or 3, **characterized in that** an individual speaker adapted model set is adaptated on basis of hyper parameters of the corresponding speaker.

5. Method according to anyone of claims 1 to 4, **characterized in that** the speech recognition is performed on basis of Hidden Markov Models.

6. Recognition system, comprising

a microphone (1) to receive spoken words of a user and to output an analog signal;
an A/D conversion stage (2) connected to said microphone (1) to convert said analog signal into a digital signal;
a feature extraction module (3) connected to said A/D conversion stage (2) to extract feature vectors of said received words of the user from said digital signal;
a recognition module (5) connected to said feature extraction module (3) to recognize said received words of the user on basis of said feature vectors;
an adaptation module (6) receiving the recognition result from said recognition module (5) to generate and/or adaptate a speaker adapted model set;

characterized by

a verification module (4) identifying a new speaker and selecting an individual speaker adapted model set that forms the basis for speech recognition of said identified speaker and model adaptation to said identified speaker.

7. Recognition system according to claim 6, **characterized by** a storage (7, 8, 9, 10) for a speaker independent model set and each individual speaker adapted model set including the adaptation hyperparameters.

8. Recognition system according to claim 7, **characterized in that** respective adaptation hyper parameters are stored in a storage (8, 9, 10) of a corresponding individual speaker adapted model set.

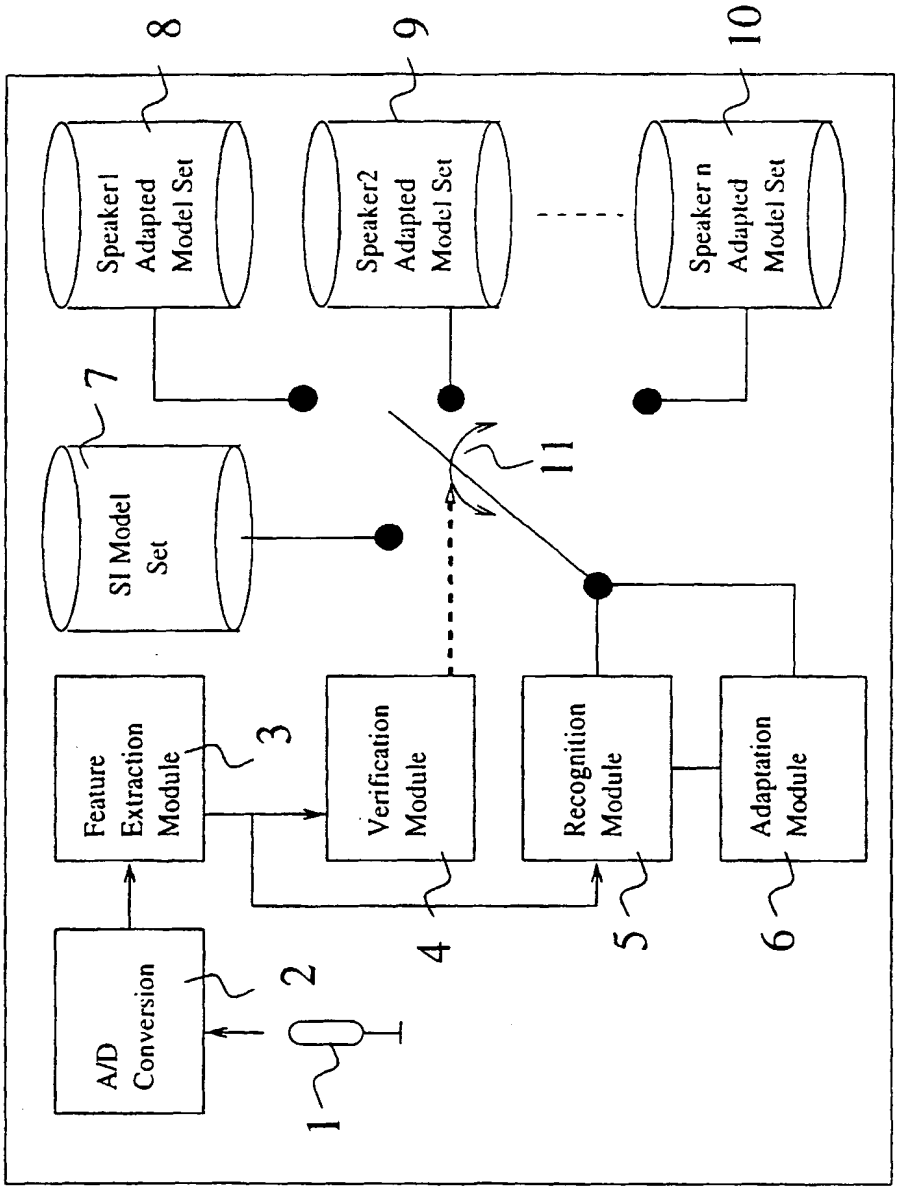

Fig. 1

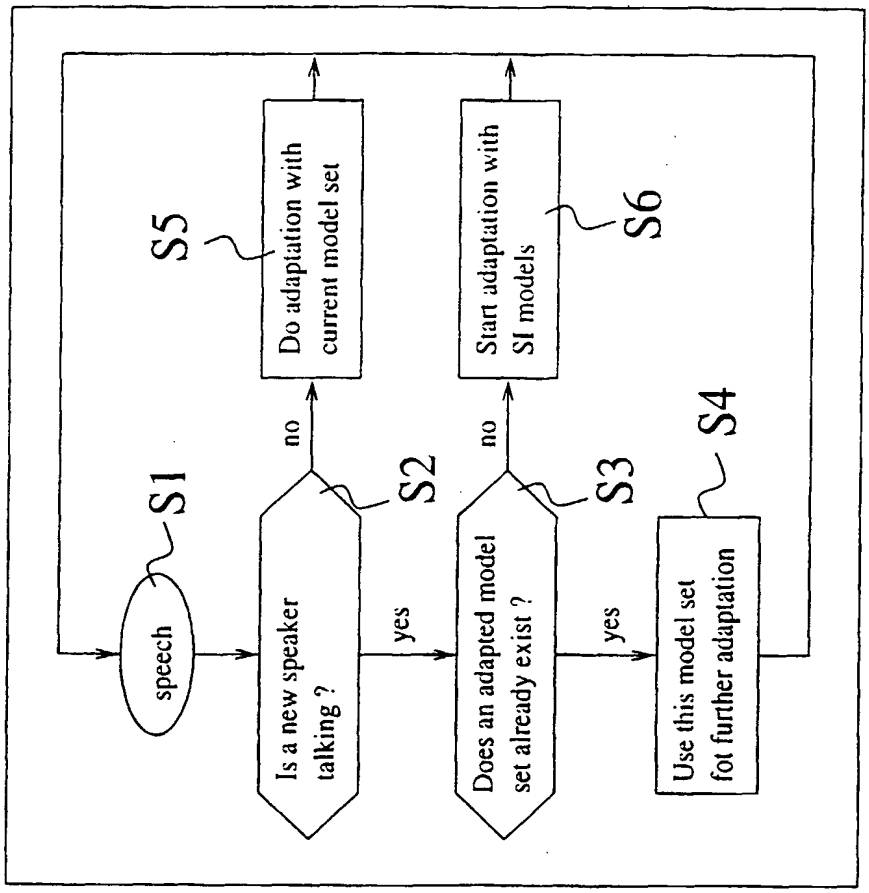

Fig. 2

European Patent Office

EUROPEAN SEARCH REPORT

Application Number
EP 99 10 0951

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION
X	REYNOLDS ET AL.: "Integration of speaker and speech recognition systems" INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING (ICASSP 1991). vol. 2, 14 - 17 May 1991, pages 869-872, XP000222216 TORONTO, CA ISBN: 0-7803-0003-3 * paragraph '0001! *	1,2,6,7	G10L17/00 G10L15/06
A	WO 96 22514 A (SRI INTERNATIONAL) 25 July 1996 (1996-07-25) * page 3, line 24 - page 4, line 2 *	1,6	

TECHNICAL FIELDS SEARCHED
G10L

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
THE HAGUE	7 July 1999	Lange, J

CATEGORY OF CITED DOCUMENTS
X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons
& : member of the same patent family, corresponding document

EPO FORM 1503 03.82 (P04C01)

ANNEX TO THE EUROPEAN SEARCH REPORT ON EUROPEAN PATENT APPLICATION NO. EP 99 10 0951

This annex lists the patent family members relating to the patent documents cited in the above-mentioned European search report. The members are as contained in the European Patent Office EDP file on. The European Patent Office is in no way liable for these particulars which are merely given for the purpose of information.

07-07-1999

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 9622514 A	25-07-1996	US 5864810 A	26-01-1999
		CA 2210887 A	25-07-1996
		EP 0804721 A	05-11-1997
		JP 10512686 T	02-12-1998